Veridicality Annotations in Spanish

by

Teresa Rosa Martín Soeder

Master Thesis

Chair
Faculty of Philosophy
Department of Language Science and Technology
Saarland University

Supervisor
Prof. Vera Demberg

Advisor
Dr. Lucia Donatelli

Reviewers
Who will be grading your thesis
The second person grading your thesis

May 20, 2023

Declaration of Authorship

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides Statt, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Statement in Lieu of an Oath

I hereby confirm that I have written this thesis on my own and that I have not used any other media or materials than the ones referred to in this thesis.

Einverständniserklärung

Ich bin damit einverstanden, dass meine (bestandene) Arbeit in beiden Versionen in die Bibliothek der Informatik aufgenommen und damit veröffentlicht wird.

Declaration of Consent

I agree to make both versions of my thesis (with a passing grade) accessible to the public by having them added to the library of the Computer Science Department.

Datum/Date: _____

Unterschrift/Signature: _____

SAARLAND UNIVERSITY
Chair
Department of Language Science and Technology

Abstract

Veridicality Annotations in Spanish

by Teresa Rosa Martín Soeder

In veridicality studies, an area of research of Natural Language Inference (NLI), the factuality of different contexts is evaluated. Here, we aim to reduce the lack of research in this field, particularly in Spanish, by presenting the analysis of annotations collected in a pilot study for two different contexts, mood alternation and specificity. Results show a inter-annotator agreement score of $AC_2 = 0.484$, slightly lower than that of de Marneffe et al. [1] ($\kappa = 0.53$), a main reference to this work. Furthermore, a significant effect of some mood alternation conditions, and a few unexpected tendencies, like high factuallity despite the presence of a negation marker, were found. Finally, suggestions that could explain the lack of inter-annotator agreement as well as improve it for the final study are defined. The annotations collected are available on `https://github.com/narhim/veridicality_spanish`.

Acknowledgements

Thx Mom+Dad < 3

Content

Dedicated to WHATEVER.

Chapter 1

Introduction

1.1 Introduction

Is that true? Did it really happen? Often we ask ourselves or our interlocutors these questions, we try to assess if the information conveyed is likely to be truthful or not, that is, if it is presumably a fact or not. Also, as speakers or authors we usually try to portray what we know about the truthfulness of the events conveyed. But, how do we do these operations? There are different linguistic factors at play, like the use of negation or modality markers, and their study is what is called veridicality, research topic for this thesis.

To be more specific, veridicality is an area of research within natural language inference (NLI) and theoretical linguistics that studies the truth value of a proposition or event in a specific context [2, 3]. As to NLI, it is a branch of natural language understanding (NLU) with its main task being entailment classification, that is, given a premise like *He believes her father has arrived*, and a hypothesis like *Her father hasn't arrived*, the task is to classify the relationship betweem then by picking a label from a usually small set of labels like {*entailment, neutral, contradiction*} [4] or {*yes, unknown, maybe*}, depending on how the task is defined. For a system to successfully complete this task, it needs to form a thorough and complete meanig representation of both sentences [4], and here is where veridicality, among other disciplines, comes into play.

The focus of thesis is the analysis of veridicality judgments in Spanish; that is, the analysis of factuality judgments in specific contexts in Spanish. In particularly, the goal is to analyze how mood alternation, in other words, the possibility of using a verb either in indicative or subjunctive mood; and the specificity of the syntactic subject affect factuality judgments. This goal is realized in the following research questions:

RQ1.- In a complex sentence, how does the mood alternation of the embedded verb that occurs due to the negation of the main or matrix verb affect the factuality value of the embedded event?

RQ2.- In a simple sentence, how does the mood alternation caused by an adverb of doubt or possibility affect the factuality judgment of the event?

RQ3.- How does an individual subject affect the factuality judgment of the event?

RQ4.- How does a subject that refers to a collective entity like an institution, affect the factuality judgment of the event?

In order to answer these questions and verify their relevance, I run a pilot study in which annotations for a small corpus were gathered. Based on the results of this study, introduced in Chapter 3, some modifications were made for the final study, the most important being the removal of research question RQ2.-. Thus, for the main or final study the research questions are the following:

RQ1.- In a complex sentence, how does the mood alternation of the embedded verb that occurs due to the negation of the main or matrix verb affect the factuality value of the embedded event?

RQ2.- In a simple sentence, how does the mood alternation caused by an adverb of doubt or possibility affect the factuality judgment of the event?

RQ3.- How does an individual subject affect the factuality judgment of the event?

RQ4.- How does a subject that refers to a collective entity like an institution, affect the factuality judgment of the event?

The main motivation behind this thesis is to try to reduce the lack of research in veridicality, specially in Spanish, that I have encounter. To my knowledge, most of the research in veridicality, in any language, is often either done from a purely thereotical perspective or completely absorved by the term factuality, thus making it difficult to find research in veridicality from a natural language processing (NLP) perspective. Furthermore, as is often the case in any NLP area, it seems that most of the datasets available for NLI are in English, thus carefully creating and making available an annotated dataset in Spanish could help better understand veridicality contexts in Spanish and also further develop NLI systems for Spanish, and maybe even related languages.

Next, Chapter 2 introduces the literature review and the explanation of the main concepts used here. Then, Chapter 3 presents the already mentioned pilot study and then the main study is introduced in Chapter **??**. Finally, Chapter **??** brings together all the issues discussed in each experiment, suggests possible improvements, and proposes some lines of future work.

Chapter 2

Background

The goal of this chapter is to clarify the most important terms that will be used in this thesis, give theoretical support for the claims made for the experiments, and present some previous research important to the study here. Specially the focus is on distinguishing concepts that are often confused, like veridicality and factuality, and introducing the reader to the phenomena in Spanish linguistics on which this thesis focuses.

First Section 2.1 clarifies the difference between veridicality and factuality, provides specific work on each of them, and explains the two main approaches used to study these concepts. Secondly, in Section 2.2, the first experimental condition in this thesis will briefly explained, mood in Spanish, together with the specific mood contexts set as experimental conditions. After this, in Section 2.3 the second experimental condition, specificity, is introduced, together with the particular specificity contexts whose veridicality is here analyzed.

2.1 Veridicality and Factuality

Let us consider examples (1a) to (1d), where we have events that are intrinsically related. If we were to do a NLI study with these examples, we could directly study the *thruthfulness* of each of them as a single event, i.e., we could study the **factual nature** of each example towards the real world or the events in the discourse [5]. Another option would be to study the factual nature of the event *Anna's father has arrived* in the different contexts in which is presented: standing completely on its own (1a), or as part of a complex event (1b) to (1d). In this case, the goal would be not to understand the factuality of *Anna's father has arrived*, but rather to understand how its factuality changes when the event is embedded under an epistemic verb (1b), a verb of believe

(1c), and a verb of speech (1d). This would be considered a **veridicality study**, and the former case, a factuality study.

(1) a. Anna's father has arrived.
b. John knows that Anna's father has arrived.
c. John believes that Anna's father has arrived.
d. John says that Anna's father has arrived.

Probably one of the most important studies in NLI with a factuality focus is that of Williams et al. [4]. They extended the work of Bowman et al. [6], which presented a NLI corpus where the premises were crowdsourced figure captions and hypothesis were also crowdsourced, by increasing the number of genre to a total of 10: transcribed conversations, official documents, letters, the public report of 9/11, non-fiction works, popular cultural articles, telephone transcriptions, travel guides, short posts about linguistics, and fiction works. This corpus, denominated as MNLI or Multi-NLI, uses the labels {*entailment, unknown, contradiction*} and is now an important benchmark in NLI, as proven, for example, by its use to test BERT [7]. From this corpus originates the XNLI corpus [8], which consists on translations of the MLNI corpus into different lanaguages, including Spanish. As we will see on Chapter **??**, a small subset of this corpus was used in the main study.

Another important work in NLI is the FactBank corpus [5], which consists on $9,488$ manually annotated events by experts which resulted in the FactBank corpus. It can be said that it is done from a factuality focus, since it does not include a systematic analysis of different contexts, but it is true that the authors present an extensive analysis of the different factors that affect the factuality of an event, that is, they present a theoretical veridicality analysis. Another important feature of their work is their set of labels, which is the baseline of the one used here, and which results from the combination of an epistemic scale, {*certain, likely (probable), possible*}, with a quality scale, {*positive, negative*}. They map this combination to the traditional Square of Opposition, but for the sake of simplicity, here this square is reduced to a linear scale of factuality, as we will see later on. In any case, the labels they used are the following: *certainly yes* (CT+), *probably yes* (PR+), *possibly yes* (PS+), *certainly not* (CT-), *probably not* (PR-), *possibly not* (PS-), *unknown or uncommitted* (Uu), and *certain but unknown output* (CTu).

An interesting factuality study is that of Pavlick and Kwiatkowski [9]. Their goal was to determine whether the disagreement often seen in NLI datasets is noise or a reproducible signal, and thus it should be included in data analysis and modelling. To fulfill this goal, they collected factuality judgments on 500 pairs from different corpora and with 50 annotators per pair. But after filteering the annotations, 496 pairs with a mean of 39 workers per pair were left to analyze.

When analyzing the annotations, the authors assumed that if the disagreement is noise, the gathered labels can be modeled as a simple Gaussian distribution where the mean is the true label and, consequently, there is only one true label. To verify this assumption, the authors fixed two models for each pair: one single Gaussian and a Gaussian Mixture Model (GMM) where the number of components is chosen during training. Results showed that overall there was a better fit with GMMs and that for 20% of the pairs there was a nontrivial second component, that is, that for 20% of the pairs there is not just one *true* label, but rather two. This phenomenon can be referred to as **label split** [1] and might be helpful in explaining the results obtained in the main study here.

Furthermore, Pavlick and Kwiatkowski [9] analyzed whether context reduces disagreement by collecting annotations at three levels: word, sentence, and paragraph. Results showed that disagreement increases with context, which the authors interpret as first evidence that agreement does not improve with more context. They explain this by hypothesizing that less context may result in higher agreement because with less context humans can more easily call upon *default* interpretations. Although context was not included in either of my studies, the work of Pavlick and Kwiatkowski [9] will help on the discussion of the results.

A clear example of a study on veridicality is that of Ross and Pavlick [10]. Their goal was to learn whether neural models with no explicit knowledge of verbs' lexical categories can make inferences about veridicality consistent with human inferences. To do so, they crowdsourced human judgments on 1,500 sentences with 137 verb complement constructions using as labels a 5-point Likert scale, and compared these judgments with those made by BERT. Their results and analysis are quite thorough, and led them to conclude that although BERT was able to replicate many of the human judgments, there is still significant room for improvement. Important from this work is also their explanation on the different perspectives that veridicality and factuality studies can have: lexical semantics or sentence meaning approach, and pragmatic or speaker meaning approach. But before explaining their definition of these approaches, we should go over the work

of de Marneffe et al. [1].

The study of de Marneffe et al. [1] is a main reference to this thesis. Their goal was to identify some of the linguistic and contextual factors that shape reader's veridicality judgments. To do so, they present crowdsourced annotations on a part of the FactBank corpus with the same set of labels minus one, CTu, and a system for veridicality assesment. For our purposes, the two most important parts of their work is the consideration of the possible occurrence of label split amd the comparison they made between their annotations and the original FactBank's annotations.

In order to see if there is a possibility of label split, they studied the plotted distribution of agreement patterns. After discarding the likely noisy patterns, they observed that they are many examples for which is unlikely that the agreement pattern is due to noise. For example, for the premise *In a statement, the White House said it would do "whatever is necessary" to ensure compliance with the sanctions.*, the judgment depends heavily on the speaker's previous knowledge about the White House.

The comparison they make between the two set of annotations is very important because it shows quantitative differences between what they called, a lexical approach and a pragmatic approach. They do not define in detail the **lexical approach**, but they do mention that in such approach the context in which we study the factuality of a proposition is a lexical item. Ross and Pavlick [10] extends this definition by adding that a system following it should aim to model the aspects of a sentence semantics, thus, this representation can be derived from the lexicon and is independent of context, and also, as stated in Saurí and Pustejovsky [5], such a system would no consider world knowledge. Thus, linguistic experts are usually the ones who annotate the data.

As to the **pragmatic approach**, approach used by de Marneffe et al. [1], by Ross and Pavlick [10], and here, it requires the system to derive a representation of the sentence that considers the communication intent for that sentence in a specific context, that is, a goal-directed representation of a sentence within the context it was created [10]. Such a representation entails two important things: the consideration of world knowledge, and the embracing of uncertainty [1]. To obtain this representation a key step is to collect annotations from linguistically naive workers, so for examples (1a) to (1d) they would consider any knowledge they might have about John, Anna, and her father; and they would ignore any notions as to what *to say* is *supposed* to mean, and instead just use

their linguistic intuition.

There are other two important facts to consider about these two approaches. The first one is that not everyone explicitly defines the approach they use, but it is often, if not always, possible to infer it. Considering this is very important, specially when comparing results from different studies. The second fact is how these approaches are referred to. de Marneffe et al. [1] and Ross and Pavlick [10] use the terms lexical and pragmatic, but Ross and Pavlick [10] also employs the terms sentence meaning approach and speaker meaning approach respectively. Furthermore, de Marneffe et al. [1] hints at the idea that the pragmatic approach is done from the reader's persective, since we want to analyze what the reader understands, not what the author says, as in the lexical approach [5]. Some studies and tasks like FACT at the Iberian Languages Evaluation Forum (IberLEF) [11], make use of these terms, author's or reader's perspective, rather than semantic or pragmatic approach.

Now that we have seen the differences between a factuality and veridicality focus, and between a lexical and a pragmatic approach, we can go on to exploring another important topic in this thesis: mood in Spanish.

2.2 Mood in Spanish

Simply put, **mood** is the grammaticalization of modality [12, 13], and thus it has been traditionally related with the speaker's attitude towards an utterance [12, 14]. But, as stated in Real Academia Española [14], this notion is imprecise and more needs to be said in order to explain all the phenomena. Nevertheless, to avoid complicated thereotical discussions, here mood is just considered as the gramatical category that reflects the commitment of a speaker towards an utterance [14].

Since the commitment of the speaker usually takes form in different degrees [12], in most languages mood takes form in different subcategories, like the indicative, the subjunctive and the imperative in Italian, or the subjunctive and the conditional in Hungarian. For Spanish, as in Italian, nowadays most of the grammarians agree on the existence of three subcategories of mood[1]: indicative, subjunctive and imperative. Since the imperative mood is out of the scope of this thesis, I would simply say that it is the mood mainly

[1]For a diachronic review of the study of mood in Spanish see Calvo [15]

used to express commands. As to the indicative and subjunctive mood, the topic is not so clear.

Quite often, the indicative and subjunctive moods are defined in opposition to each other Lyons [12], and Spanish is not and exception. Some pairs of concepts that Real Academia Española [14] uses to describe this opposition are the following: certainty/uncertainty, reality/virtuality, actuality/non actuality and commitment of the speaker with the veracity of what is spoken/lack of commitment. So for example, in (2a), where the verb *estar* (to be) is in indicative, the reading that we get is that *it is a reality that Sofía was there to see it.* Contrary to this, in (2b), where the same verb, *estar*, is in its past perfect tense from the subjunctive, the reading is that *Sofía wasn't there, but we wish there was world, a possible world, in which she was there.* But, as stated in Real Academia Española [14] and Sánchez-Jiménez [13], these oppositions do not always work well.

(2) a. Sofía **estuvo** allí para verlo.
Sofía **be.pst.pfv.ind.3sg** there to see.INF.it
"Sofía was there to see it."

b. ¡Si Sofía **hubiera estado** allí para verlo!
if Sofía **have.pst.pfv.sbjv.3sg be.ptcp** there to see.INF.it
"If only Sofía had been there to see it!"

(3) a. Quiero suponer que **has preparado**
want.PRS.IND.1SG assume.INF that **have.prs.ind.2sg prepare.ptcp**
todo.
everything
"I want to assume that you **have prepared** everything."

b. Siento mucho que se te
feel.PRS.IND.1SG a.lot that itself you.DAT
haya averiado el coche.
have.prs.sbj.3sg break.down.ptcp the.M.SG car
"I'm so sorry to hear that you car **broke down**."

For example, in (3a), where the embedded verb *preparar* (to prepare) is indicative, the reading for the event is *I want the fact that you have prepared everything to be a reality, but it might not be*; and, opposite to this, in (3b), where the embedded verb *averiarse* (to break down) is in the subjunctive mood, its reading is that *it is a reality that your car broke down and I feel sorry for that.* Does this mean that the above mention oppositions are useless? No, but it does imply that interpreting them too strictly would be a mistake, and that some complimentary considerations are necessary. In this regard, Villalta

[16] talks about an ordering relation between contextual alternative propositions as the cause for embedded propositions to be in the subjunctive mood, and Mejías-Bikandi [17] explains the contrast in terms of old and new information.

Specifically, Mejías-Bikandi [17] understood **old information** as the information that is pragmatically presupposed, or in other words, the information with which the speaker assumes familiarity. **New information** would be then the one that is not presupposed. Based on these definitions, he classifies **matrix verbs**, that is, verbs that take a verbal predicate as a complement, into two groups: those that introduce old information and those that introduce new information. Then he uses this classification to explain the distribution in mood in complements of different types of matrices. The indicative, he claims, is used when the matrice introduces new information, as in the case of mental matrices like *notar* (to notice); and the subjunctive is instead used when the matrice introduces old information, as in the case of comment matrices like *lamentar* (to regret). Furthermore, he uses this classification to explain the distribution of mood for the following phenomena: the negation of some matrix verbs and matrix verbs whose subject contain a quantifying expression such as *poco/a/s* (a bit/bits) or the adverb *solo* (only). Next, we will briefly go over the syntactic functioning of the indicative and subjunctive moods.

As stated in Real Academia Española [14], authors often talk about a **dependent** and an **independent mood**, the first one being the one that requires a grammatical inductor in order to appear, like in example (2b), where the conditional conjunction *si* forces the subjunctive in the verb *estar*; and the second being the one that does not need any grammatical element in order to appear in the sentence. This distinction mostly correlates with the subjunctive and the indicative moods, since the cases in which the subjunctive appears without depending on any inductor are highly restricted [14], and the preferred choice for most of the simple sentences is the indicative, even if in many induced contexts choosing the indicative mood is a requirement, as in example (3a).

An important characteristic of these induced contexts, is that the specific mood, let it be the indicative or the subjunctive, prompted by the grammatical inductor can be an imposition, like in examples (2b) and (3b), or a choice, as in examples (4a) and (4b), and this is what is called **mood alternation**. This last case is the main feature of the experimental analysis in this thesis, and as such, it will be explained in detail next.

2.2.1 Mood Alternation

As stated above, mood alternation is the phenomenon in which a particular mood is induced within a distinct structure, usually the subjunctive in nominal complements, but this induced mood is *optional*, that is, speakers use the structure with either the indicative or the subjunctive mood, as in the examples below. There we can see that the two sentences are almost identical, even the translations into English are the same[2], the only visible difference lays on the morphology of the embedded verb. *Tener* (to have) is in the indicative mood in (4a), an in the subjunctive mood in (4b). What causes this phenomenon? How is it interpreted by speakers?

(4) a. El presidente no dijo que el
the.M.SG president.M.SG not say.PST.PFV.IND.3SG that the.M.SG
país **tenía** problemas económicos.
country.M.SG **have.pst.ipfv.ind.3sg** problem.M.PL economic.M.PL
"The president didn't say that the country **had** economic problems."

b. El presidente no dijo que el
the.M.SG president.M.SG not say.PST.PFV.IND.3SG that the.M.SG
país **tuviera** problemas económicos.
country.M.SG **have.pst.ipfv.sbjv.3sg** problem.M.PL economic.M.PL
"The president didn't say that the country had economic problems."

Contrary to what we have seen in the previous section about the overall distinction between indicative and subjunctive, in mood alternation the difference is quite clear. As stated in Real Academia Española [14], Mejías-Bikandi [17] and Falk and Martin [19], the indicative is used when the speaker chooses to present the event as new information, and the subjunctive mood is used when the information is already part of the common ground. Thus, in example (4a), the fact that *the country has economic problems* is presented as something new to the speaker; and in (4b), the fact that *the country has economic problems* or that *the country doesn't have economic problems* is considered to be known by the speaker.

An important work on mood alternation is the study of Faulkner [18]. Her goals were to see if there are non-standard cases of mood alternation in verbal complements and if these acceptability is dialectal, to understand if the informativeness of the complements affects this acceptability, and if so, to see in which cases this happens. To fulfill these

[2]There are ways in which the difference between the indicative and the subjunctive can be translated into English, but given that they are not very obvious or common translations, I chose to make no distinction, as in Faulkner [18].

goals she collected acceptability judgments on 128 sentences, with and without context, of speakers from different Spanish varieties. For all the sentences the induced mood is the subjunctive and the alternative is the indicative.

The main idea from the results of this study is that mood alternation in verbal complements is a complex phenomenon in which different factors come into play. First of all, the type of complement, or in other words, the type of matrix verb, considerably affects the acceptability judgment. For example, desideratives don't accept the indicative in their complements, but negated epistemics do. Secondly, moods alternation seems to depend on the presence or absence and on whether the context is informative or not. Thirdly, she shows some dialectal variations which we will neither discuss nor forget here. Lastly, it should be noted that, despite the thorough analysis she presents, not all results could be explained, which shows that further research into mood alternation is needed. Next the specific mood alternation phenomena analysed in this thesis are presented.

The first type is the negation of the matrix verb, shown in (4a) and (4b). As stated in Real Academia Española [14], negation can induce the subjunctive in the verb of the nominal complement. Specifically, they say that the adverb *no* (no) induces the subjunctive mood in the complements of verbs of speech, perception, thought and believe; and they give examples of mood alternation for verbs of perception. Thus, for the pilot study, only these categories of matrix verbs were considered.

The second type of mood alternation phenomena is that of adverbs of doubt and possibility, exemplified in (5a) to (5d). As stated in Real Academia Española [14], adverbs of doubt and possibility such as quizás (maybe) or *probablemente* (probably), induce indicative or subjunctive within their own sentences, but the subjunctive mood appears only when the adverb precedes the verb and there is no pause between them, as in examples (5a) and (5b). To my knowledge, research on this topic is not abundant, thus I was not able to define what exactly that *pause* means, that is, if it is literally a pause like the one indicated by *,* and other elements can be in between, such as the subject as in examples (5c) and (5d); or if it means that the adverb must inmediately precede the verb and that no element can be in between, as in examples (5a) and (5b). For the pilot study, the former option was assumed to be true, but knowing that it could *easily* be denied. Next the second feature analyzed in the pilot study is presented.

(5) a. Quizás **dijo la verdad**.
maybe **tell.pst.pfv.ind.3sg** the.F.SG truth.F.SG
"Maybe s/he **told** the truth."

b. Quizás **dijera** la verdad.
maybe **tell.pst.pfv.sbjv.3sg** the.F.SG truth.F.SG
"Maybe s/he **told** the truth."

c. Quizás el presidente **dijo** la verdad.
maybe the.M.SG president.M.SG **tell.pst.pfv.ind.3sg** the.F.SG truth.F.SG
"Maybe the president **told** the truth."

d. Quizás el presidente **dijera** la verdad.
maybe the.M.SG president.M.SG **tell.pst.pfv.sbjv.3sg** the.F.SG truth.F.SG
"Maybe the president **told** the truth."

2.3 Specificity

If we consider the dialog below we can see a very simple interaction in which *A* informs *B* about the location of a specific book. But that is not everything that is being communicated. If we look at the utterance that *A* produces, we can see that aside from the location of an entity, *A* is making reference to two entites: a specific book and a specific table. But what does referring to something mean?

(6) A. *The book is on the table.*
B. *Ok. Thank you.*

To **refer** to something is to point, pick up, or call up on an entity or set of entities [12] in the mind of the speaker [20]. In other words, when a speaker makes a reference, she presupposes the existence of an entity or set of entities, and connects a linguistic expression to it [21]. When the listener hears such expression, he tries to make that same connection by correborating its existence in the physical context, linguistic context, and/or his memory. This is why, the study of referential expresions is intrinsically connected to existence [12], to the study of existential presuppositions; to the point that, as stated in García Murga [21], the study of existential presuppositions must be supported by the study of reference. Further clarifications about this relation can be seen in Lyons [12], García Murga [21] and Herrasti et al. [22].

Above we have mentioned that the book and table to which *A* refers are a specific book and a specific table. At first glance, this can be simply understood as a particular book and a particular table, but as Caudet [20] thoroughly explains, the property

of being specific has being applied to different concepts, which can lead to some confusion. Thus she puts together the different views on specificity and summarizes them into the following six criteria, criteria that can be used to classify referential expressions:

1. Existence of the referent in the extra-linguistic reality.
2. Identifiability of the referent in the extra-linguistic reality.
3. Extension indicated by the referential expression.
4. Existence of the referent in the discourse universe.
5. Identifiability of the referent in the discourse universe.
6. The expresion points to a set of referents pragmatically delimited or chooses a subset of them.

Here we consider **specificity** in the sense of the fifth criterion, therefore when manipulating the specificity of a nominal phrase, we indicate that the identifiability of the referent in the discourse universe is what is being analyzed. In particular, here we manipulate it by modifying the amount and type of information given in the subject of the premise.

Regarding the type of information, two manipulations are performed: individual vs. collective nouns, and common vs. proper nouns. With the first one we are manipulating the amount of individual entities to which we are referring to. By using an individual noun (in singular) like *president*, we point to one single entity, whereas by using a collective noun like *government*, we point instead to a set of entities; thus when trying to identify the referent in the discourse universe, the set of possible referents the listener considers is different.

As to the second manipulation, common vs. proper nouns, we alternate who identifies the referent. When using a common noun like *president*, the speaker only identifies the set of possible entities to which our referent belongs to, therefore leaving to the listener the final determination of the particular referent. Contrary to this, when using a proper noun, the speaker *usually*[3] points directly to an entity of set of entities. Therefore, by distinguishing between common and proper nouns we have two different ways

[3]For some exceptions on the common use of proper nouns see Caudet [20].

of identifying the referent: identification by the listener and identification by the speaker.

With respect to the amount of information, only one manipulation is considered: is the common noun in the subject modified only by a determiner as in *the president*, or are there any other complements that modifiy the noun as in *the president of the government*? By performing this manipulation, which here we denominate mixed, we are influencing the retrieval of the referent in a similar way as to the common vs. proper noun distinction since we are modifying the number of possible referents the listener considers. When we have a mixed noun phrase, this number is smaller than with an almost bare noun phrase, but bigger than with a proper noun, thus being a sort of middle ground between common and proper nouns and resulting in the following scale of number of possible referents: common > mixed > proper.

Chapter 3

Pilot Study

3.1 Design

The goal of this pilot study is three-folded. First, I wanted to crowdsource annotations on premise-hypothesis pairs to answer the research questions introduce in Chapter 1 and which are repeated below:

RQ1.- In a complex sentence like examples (4a) and (4b), how does the mood alternation of the embedded verb that occurs due to the negation of the main or matrix verb affect the factuality value of the embedded event?

RQ2.- In a simple sentence like examples (5c) and (5d), how does the mood alternation caused by an adverb of doubt or possibility affect the factuality judgment of the event?

RQ3.- How does an individual subject, like the one in example (5c), affect the factuality judgment of the event?

RQ4.- How does a subject that refers to a collective entity like *the government*, affect the factuality judgment of the event?

These questions yield the four experimental conditions: negation (RQ1.-), adverb or possibility (RQ2.-), individual (RQ3.-), and collective (RQ4.-). Furthermore, given the novelty of these conditions and my unfamiliarity with a study of these dimensions, the second and third goals of the study are to ensure the correctness of the experimental design and procedure, and to explore different possibilities of statistical analysis, so that in the final study the data is properly analyzed and represented.

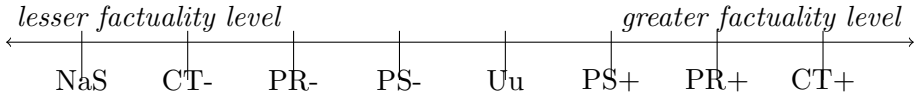

FIGURE 3.1: Ordered representation of the labels used for annotating the corpus. Each label stands for: certainly yes (CT+), probably yes (PR+), possibly yes (PS+), unknown or uncommitted (Uu), possibly not (PS-), probably not (PR-), certainly not (CT-), not a sentence (NaS).

If we take the four conditions, sort them into two groups, mood alternation for negation and adverb, specificity for individual and collective; and cross them, we obtain a 2x2 design, where each conditon is divided into 3 subconditions or categories. The negation and adverb conditions are divided into three categories: baseline (B), where there is no negation of the matrix verb or presence of a doubt or possibility adverb, like in *El presidente dijo la verdad* (The president told the truth); indicative (I), where we have a mood induced adverb and the verb is in indicative mood like in (5c); and subjunctive (S), where we have a mood induced adverb and the verb is in subjunctive mood, as in (5d).

The specificity conditions are also divided into three categories: common (C), where the subject is a common noun with the determiner as its only complement as in *the president*; mixed (M), where the subject is a common noun but with a determiner and another complement like in *the Spanish president*; and proper (P), where the subject is directly a proper noun like *José Luís Rodríguez Zapatero* or there is a relation to a proper noun, like in *Ronaldinho's mother*.

As to the set of labels used to annotate each pair, I follow de Marneffe et al. [1] and used the labels from Saurí and Pustejovsky [5] minus *certain but unknown output* (CTu). But since, as mentioned in the previous chapter, I was not certain about the acceptability of sentences like (5c), where we have a nominal phrase between the adverb of possibility and the verb, I added another label, *not a sentence* (NaS), resulting in the following set of 8 labels, represented for simplicity as a one-dimension scale in Figure 3.1: certainly yes (CT+), probably yes (PR+), possibly yes (PS+), unknown or uncommitted (Uu), possibly not (PS-), probably not (PR-), certainly not (CT-), not a sentence (NaS).

3.2 Dataset

As to the dataset to be annotated, it was constructed by first writing 9 seed premises for each of the four baseline-common combinations and from them the hypotheses were extracted to form the pairs. Then these seeds were changed across conditions, yielding a total of 324 premise-hypothesis pairs[1]. Table 3.1 shows this generation process together with the experimental design.

Data Generation within Experimental Design								
			SPECIFICITY					
			Individual			Collective		
			C	M	P	C	M	P
MOOD	Negation	B	S1	V1	V1	S2	V2	V2
		I	V1	V1	V1	V2	V2	V2
		S	V1	V1	V1	V2	V2	V2
	Possibility	B	S3	V3	V3	S4	V4	V4
		I	V3	V3	V3	V4	V4	V4
		S	V3	V3	V3*	V4	V4	V4

TABLE 3.1: Experimental conditions and corpus generation. Each cell is for 9 pairs. **S** followed by a number stands for set of seeds, **V** stands for variants, with the number indicating the corresponding set of seeds. ***** indicates where the 2 mistaken pairs are. **B, I, S** stands for *baseline, indicative, subjunctive*, and **C, M, P** stands for *common, mixed, proper*.

3.3 Predictions

Now that we the different variables considered in the study have been explained, we can try to envision how they will affect the annotations. As we can see in Table 3.2, the highest factuality labels are predicted for the baseline category, since in the absence of any of the mood inducing conditions, no other factuality modifier is expected, for the exception of the specificity categories. When combined with them, the factuality level of the baseline pairs might change, but only within the range specified in each case. This is because by modifying the amount of information (mixed category) or the type of information (proper category), it is expected that world knowledge, and thus uncertainty, plays a greater rol in assigning the label.

As to the other two mood categories, indicate and subjunctive, a decrease on the factuality level with respect to the baseline is to be expected, with the indicative having

[1]Upon reviewing the pairs after the experiment was run, it was noticed that there was a mistake on 2 pairs, thus yielding them invalid. The whole two seeds were taken out for analysis, yielding a total of 306 analysed pairs.

Label Predictions within Experimental Design								
			SPECIFICITY					
			Individual			Collective		
			C	M	P	C	M	P
MOOD	Negation	B	CT+	CT+-PR+	CT+-PS+	CT+	CT+-PR+	CT+-PS+
		I	PR+	PR+-PS+	PR+-Uu	PR+	PR+-PS+	PR+-Uu
		S	PS-	PS+-PS-	PS+-PS-	PS-	PS+-PS-	PS+-PS-
	Possibility	B	CT+	CT+-PR+	CT+-PS+	CT+	CT+-PR+	CT+-PS+
		I	PR+	PR+-PS+	PR+-Uu	PR+	PR+-PS+	PR+-Uu
		S	PS+	PS+-Uu	PS+-Uu	PS+	PS+-Uu	PS+-Uu

TABLE 3.2: Rough approximation of label predictions for each combination of experimental categories. **B, I, S** stands for *baseline, indicative, subjunctive*, **C, M, P** stands for *common, mixed, proper*, and - indicates a range of labels when not indicating the quality of a specific label.

a higher degree of factuality than the subjunctive. Since by choosing the indicative mood the speaker presents the event, our hypothesis, as new information, she limits the possibilities of the listener belying the factuality of the hypothesis. Contrary to this, by using the subjunctive, the speaker assumes that the factuality of the event is already known and agreed upon, thus increasing the possibilities of it being denied, or in other words, she allows more uncertainty. These predicitions for the indicative and subjunctive categories roughly hold for both mood alternation conditions, with the only difference being that for the negation condition, the factuality level for the subjunctive category is predicted to be lower than for the same category in the possibility condition, reaching even the level of the negative labels. This is due to the above-mentioned fact that only in this case the negation marker scopes over the embedded event, thus reducing the factuality of the event more strongly than a mere adverb of doubt or possibility.

Regarding their variability when crossed with different specificity categories, it is expected that labels within the indicative category will decrease in a similar fashion as to labels within the baseline category; but for the subjunctive category, barely any differences are expected since values in the subjunctive category are already within the range of high uncertainty.

Lastly, it should be noted that for most of the cases the labels predicted are within the positive range of the scale. In other words, the distribution of the labels is expected to be negatively skewed.

3.4 Procedure

The annotations were collected through Google Forms. The first page of the form asked them to choose their variation of Spanish (European or American)[2], and the second one showed the instructions for the task, together with an example. It should be noted that these instructions might had been too simple, and that they should be elaborated more carefully for the final study.

After the second page, the pairs to be labelled were presented in 9 pages. Following de Marneffe et al. [1], the task consisted on questions where the labels were presented as answers in a single choice list. de Marneffe et al. [1] presented the labels ordered, but, by mistake, in this study labels were not completely ordered.

Another important feature of this experimental procedure is that, since cognitive overload and an overly time-consuming task were to be prevented, pairs were divided into 9 batches, thus having a total of 9 forms with 36 pairs each. Initially, only 3 annotators were to annotate each pair, but as the task progressed, the lack of disagreement was noticed, and therefore two more annotators were added to each batch. That is, pairs were annotated by 5 raters, and each of them annotated 36 pairs.

With respect to the annotators, they were chosen among family and friends whose mother tongue is Spanish and who do not have higher linguistic education. This last feature was important in order to prevent annotations having a lexical approach, but, some of the annotators acknowledged afterwards that, since they attempted to find the *correct* answer, they had use their linguistic knowledge from their basic education instead of their *raw* linguistic instincts; hence this was acknowledge in the instructions for the final study.

Now that we have explained the study goals, design, predictions and procedure, we can present the statistical analysis of the annotations collected.

[2]Given that the number of annotators for each group was completely different, this variable was not taken into account for the analysis.

Basic Statistics		
	Absolute	Percentage
Total number of analyzed pairs	306	100
Total number of analyzed annotations	1530	100
Number of pairs without a label	121	39.542
Number of pairs with a label	185	60.458
Agreement equal to 3	112	36.601
Agreement greater than 3	73	23.856
Maximal agreement	18	24.657
Number of pairs with at least one NaS	32	10.457
Number of times NaS label was used	37	2.418
Number of baseline pairs in which NaS was used	8	25
Number of pairs with NaS labelled as NaS	0	0
Number of problematic seeds	8	23.530

TABLE 3.3: Some basic counts of the annotations. Maximal agreement means that the 5 raters agreed on one label. A problematic seed is one than has 5 or more of its variants without a label.

3.5 Results

3.5.1 Overview of the Annotations

In order to get a first understanding of the annotations, different plots were drawn and some very basic statistics were computed. These computations are shown in Table 3.3 and the two most relevant plots are shown below. Probably the two most important things to notice are that, as predicted, in Figure 3.2, where we have the overall distribution of labels, we observe that this distribution is quite negatively skewed, despite the considerable amount of pairs with a negation marker; and that the numbers of pairs for which agreement is reached is rather low, considerable lower than in de Marneffe et al. [1], where the percentage of pairs without a label was 22.118%. In that same plot we see that the differences between the negative labels is rather small, and that the Uu and PR+ labels are used with almost identical frequency.

In Figure 3.3, where the distribution of label per batch is shown, there is further evidence of a considerable lack of agreement, since several differences between the label distributions for each batch appeared, contrary to what was expected. For example, we see that batch B clearly uses the label CT+ more often than batch E, or that batches H and I seem to use the label NaS more often than the others.

Another important observation is the use of the label NaS. This label, mainly meant to be an insurance for the possibility condition, seems to be overused. This is evident in

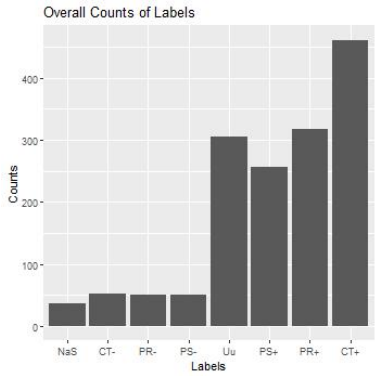

FIGURE 3.2: Overall distribution of the proportion of labels used by annotators.

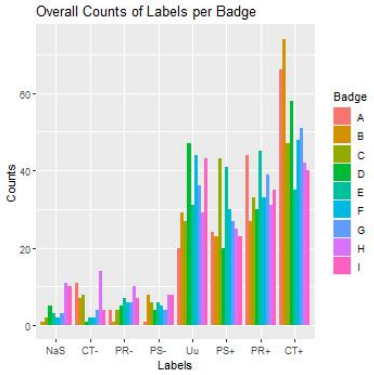

FIGURE 3.3: Distribution of the proportion of labels used by annotators grouped by batch.

the fact that although NaS was used at least once in 10.457% of the pairs, no pair was labelled as not acceptable; and furthermore, 8 pairs in which this label was used, were baseline pairs, that is, pairs without any difficulty. Therefore the role of this label must be reviewed.

To sum up, in this section we have seen that, as predicted, the distribution of the labels is negatively skewed, with most of the counts being in the positive range of the labels' scale. We have also shown that the NaS label might have been misused in the annotations. Moreover, some indications of a considerable amount of disagreement in the annotations have been presented: the number of pairs in which at least 3 annotators agreed upon a label is lower than in the case of de Marneffe et al. [1] and the labels' distributions for each batch have some clear differences. In the next section we will see the significance of this disagreement by measuring inter-annotator agreement.

3.5.2 Inter-Annotator Agreement Scores

Finding the score that correctly computes the inter-annotator agreement was a more complicated task than expected. Thus, instead of simply presenting one κ value to represent the inter-annotator agreement for the whole corpus, we present main values for different scores. All inter-annotator agreement scores computed for this study can be seen in Appendix **??**.

Agreement Scores for the Whole Dataset	
Name	Value
Fleiss' κ	0.160
Fleiss' κ with linear weights	0.177
Conger's κ with linear weights	0.176
Gwet's AC_2 with linear weights	0.484

TABLE 3.4: Inter-annotator agreement scores computed for the whole dataset.

Initially, I followed the work of de Marneffe et al. [1], and computed Fleiss' kappa [23], which resulted in a value of $\kappa^f = 0.160$, which means slight agreement according to the labels presented in Shrout [24], but since, as de Marneffe et al. [1] already remarks, this score does not consider the order between the labels shown in Figure 3.1, thus yielding an inaccurate value for this case, other possibilities had to be considered.

The main alternative is to use weighted kappa, a score that extends Cohen's kappa [25] to observations made using ordinal labels. For the case presented here, we would specifically need to extend the definition to the case of multiple raters and ordinal labels, task that is not without problems [26, 27] and thus it has different solutions. What all the different implementations have in common is defining κ as a relation between the weighted proportion of observed agreement p_{aw}, and the weighted proportion of chance agreement $p_{a|c_w}$ [28].

Here, initially the definitions presented in Warrens [29] and Warrens [30] were chosen, but after encountering the work of Vanacore and Pellegrino [28], this decision changed. Vanacore and Pellegrino [28] presents a theoretical analysis of four weighted kappas which differ in their definition of $p_{a|c_w}$, Fleiss [23] (κ_w^f), Conger [31] (κ_w^c), Brennan and Prediger [32] (s_w^*), and Gwet [33] (AC_2); as well as a comparison of their behaviours in paradoxical environments. Their results showed that both κ_w^f and κ_w^c were the most affected by these environments. So, given these results, the fact that κ_w^f and κ_w^c are, to my knowledge, more commonly used; and that, as seen in Figure 3.2, the distribution of our labels is skewed, or in other words, paradoxical; three different agreement scores were computed, κ_w^f, κ_w^c, and AC_2, with the software package R, specifically the IrrCAC library [34]. Results are shown in Table 3.4.

As we can see, all weighted scores show an improvement with respect to the unweighted Fleiss' κ, but only AC_2 clearly distinguishes itself from the kappa computed first. The other two weighted kappas are not just within the same range of sligh agreement $(0.11 - 0.40)$, but there are also barely distinguishable from the unweighted kappa. This

Agreement Scores for Experimental Conditions			
Subset	κ_w^f	κ_w^c	AC_2
ALL	0.177	0.179	0.484
Negation	0.094	0.097	0.388
Possibility	0.278	0.278	0.592
Individual	0.161	0.162	0.500
Collective	0.190	0.192	0.470

TABLE 3.5: Inter-annotator agreement scores computed for the subsets corresponding to each of the experimental conditions. Values for the whole dataset are given as reference,

could be interpreted as more evidence of very low agreement between annotators, but, given the already mentioned work of Vanacore and Pellegrino [28], that in 60.458% of the pairs at least 3 annotators agreed upon a label, and that the AC_2 has a value of 0.484, which is clearly within the range of fair agreement $(0.41 - 0.60)$; it appears rather that these results can be taken as more evidence of the sensibility of κ_w^f and κ_w^c to paradoxical behaviour. Thus it seems appropiate to consider $AC_2 = 0.484$ as the inter-annotator agreement score that represents our corpus, but in order to strengthen this decision, the three weighted scores are used in any other computations presented here. Now we will briefly explore agreement within the experimental conditions.

3.5.2.1 Inter-Annotator Agreement for each Experimental Condition

Table 3.5 shows the value of the 3 inter-annotator agreement scores computed for the subsets of the corpus that correspond to each of the experimental conditions. Again we see what we saw above, Gwet's AC_2 is clearly greater in each case, and Fleiss' and Conger's κ are roughly equal, thus strengthening the decisions to use Gwet's AC_2 score. More importantly, based on these scores we can say that the difference between the specifcity conditions is minimal, and therefore they are likely to be quite irrelevant as veridicality categories. Contrary to this, the difference between the mood alternation conditions is quite big, suggesting that they are suitable as veridicality categories. We will see afterwards if these ideas hold or not. Next we will present some simple computations done to explore the validity of the set of labels used.

3.5.3 Label Space Reductions

In order to verify the validity of our label set, or in other words, to check if the low agreement score could be partially blamed to an inadequate label set, different label

Agreement Scores for Different Label Reductions			
Reduction	κ_w^f	κ_w^c	AC_2
ALL	0.177	0.179	0.484
No NaS	0.190	0.191	0.469
PS with PR	0.186	0.187	0.572
PS with PR, no NaS	0.204	0.205	0.537
PR with CT, PS with Uu	0.153	0.155	0.635
PR with CT, PS with Uu, no NaS	0.162	0.164	0.518
PS with Uu	0.204	0.205	0.590
PS with Uu, no NaS	0.224	0.225	0.550

TABLE 3.6: Inter-annotator agreement scores computed for the different label space reductions. *No NaS* means that all pairs in which any of the annotators used the label NaS were removed. For any pair of labels linked by the preposition *with*, the interpretation is that the first label was replaced with second one, with the correspondent quality signs when required.

space reductions were performed, and the three weighted scores were computed for each of these reductions. The reductions are of two types: removal of any pair which was annotated with the label NaS, and merging of pairs of labels. By merging is understood the replacement of any instance of the first element of the pair with the second element, with the corresponding quality signs when required.

The results, shown in Table 3.6, display very interesting tendencies. First of all, the behaviours of the scores with the whole label set still holds: κ_w^f and κ_w^c are quite similar and within the range of slight agreement, and AC_2 is considerably higher and, for most of the cases, within the range of fair agreement. Secondly, we see that the scores are not always affected in the same way by the different reductions. Probably the best two examples are the *No NaS* and the *PR with CT, PS with Uu* reductions. In the first case we see that removing all pairs in which the NaS label is used benefits both κ_w^f and κ_w^c, but it has a detrimental effect on AC_2. In the second case we witness the opposite effect happening. When reducing to the typical labels {*yes, unknown, no*} (CT+, Uu, CT-), with or without NaS, it seems that it greatly benefits AC_2, but that it diminishes both κ_w^f and κ_w^c. Thus, it appears that making radical decisions about the label set might be unwise. Nevertheless, it was considered that removing PS+ and PS- could be benefitial for the final study since scores increased in every case they were removed.

Next, the results of fitting a regression model to the annotations are presented.

3.5.4 Model Fitting

In order to understand how the different variables defined in Section 3.1 influenced the annotations, a cumulative link mixed model, was fitted with a logit link by using the R software, specifically the ordinal package [35]. As predictors, mood conditions were crossed with specificity conditions and mood categories, specificity conditions were further crossed with specificity categories, and mood categories were also crossed with specificity categories. Items and annotators were set as random intercepts, and the labels as the response variable. Here I only present the most relevant results, all details are in Appendix **??**.

The most important observation is that most of the effects defined do not have a significant effect, only the mood condition adverb ($p = 2.31 \times 10^{-14}$) on its own, and crossed with the different mood categories ($p < 2 \times 10^{-16}$ for both the indicative and subjunctive category) have significant effects, wich is consistent with the results in Section 3.5.2.1. Thus we can say that the mood conditions can predict a factuality label, and that for the mood condition adverb the mood categories play a significant rol in predicting the label. As to the other predictors, there is one that is close to being significant, that of the crossed effect of the mood condition adverb with the specificity condition collective ($p = 0.054$), telling us that for the adverb condition, the specificity conditions might have a significant effect if the experimental design and procedure are improved.

The model threshold coefficients, also known as cut-points or intercepts [35] and shown in Table 3.7, represent the division of the *true* underlying factuality value for a premise-hypothesis pair. Thus they can be interpreted as the width of each category in the label set, with the exception of our two extremes of the scale. Considering this and that, aside from the PR- and PS- labels, the space for each label is roughly similar, we can say that the labels are all informative. Given this, and the fact that the computations presented in the previous section were more an exploration than a proper statistical test, the proposal to remove the PS+ and PS- labels was not considered.

To sum up, from implementing a cumulative link mixed model I learnt that mood conditions and mood categories are informative to the experimental design, although not in all settings, and that the specificity conditions could have a significant effect when combined with the adverb condition. Hence it appeared reasonable to keep these three categories of variables for the final study, but not the specificity categories. Furthermore, given the threshold coefficients obtained, the proposal to remove the PS+ and PS-

Model Threshold Coefficients			
Threshold	Estimate	Std. Error	z value
NaS—CT-	−4.254	0.326	−13.062
CT-—PR-	−3.272	0.298	−10.962
PR-—PS-	−2.756	0.291	−9.480
PS-—Uu	−2.371	0.286	−8.277
Uu—PS+	−0.904	0.278	−3.249
PS+—PR+	0.006	0.277	0.023
PR+—CT+	1.233	0.279	4.415

TABLE 3.7: Threshold coefficients resulted from fixing a cumulative link mixed model the whole dataset with different combinations of mood conditions, mood categories, specificity conditions and specficity categories; and random intercepts for raters and premise-hypothesis pairs.

labels was no longer considered. Next, we will procede to the final section of analysis of results where I will compare the annotations collected with the predictions.

3.5.5 Comparison with Predictions

Once we have obtained some numerical features of our annotations, we can dive into the data in order to examine whether the predictions were fulfilled, and what linguistic characteristics favoured or hindered agreement. To do so, I examine the labels assigned to each combination of the experimental conditions and categories presented in Table 3.8 and Table 3.9.

Assigned Labels within Experimental Design					
			SPECIFICITY		
			C	M	P
MOOD	Negation	B	CT+/PR+/Uu	CT+/PS-	PR+/Uu
		I	CT+	CT+	-
		S	?	CT+/Uu	-
	Possibility	B	CT+	CT+	CT+
		I	PR+/Uu	?	PR+/PS+
		S	PR+	-	?

TABLE 3.8: Assigned labels for the first half of each combination of experimental categories. **B, I, S** stands for *baseline, indicative, subjunctive*, and **C, M, P** stands for *common, mixed, proper* The presence of a label/s or **?** means that there were at least 5 pairs in in which 3 raters agreed upon a label, the opposite situation is indicated with **-**. **?** means there was no majority vote for any label among the different pairs. **/** means that the votes were equally divided.

There are two main remarks to make about these tables. First of all, for the indicative category there is a higher factuality level than expected, particularly when crossed with

the negation condition and to the point that, in some cases, the label assigned to the indicative pair has a higher factuality label than for its correspondent baseline pair, although the distance between these labels is, in more cases, minimal. Furthermore, heeding the considerable lack of agreement and the consequential lack of pairs with a label, its seems that this difference between indicative and baseline pairs is consistent across all variants of the respective seed. Thus we can say that this difference might be due to some lexical characteristics. Further analysis into this phenomenon was left for the final study.

The second important observation is two-folded and concerns the labels assigned to the baseline pairs in the negation condition. First of all, we see a higher disagreement than expected, indicated both by the lack of variants to set a label for the combination and by the label split, in other words, that more than one labels is assigned, for half these combinations. Second of all, we see, in some cases, a lower factuality level than expected, to the point of uncertainty (Uu label). Determining the cause of these phenomena is not straightforward, but there is one helpful fact. Upon reviewing the data, it was noticed that one factor was not taken into consideration when defining the predictions: the matrix verbs.

Assigned Labels within Experimental Design					
			SPECIFICITY		
			Collective		
			C	M	P
MOOD	Negation	B	CT+	-	-
		I	CT+	-	CT+
		S	-	-	-
	Possibility	B	CT+	CT+	CT+
		I	PR+	-	?
		S	PR+/Uu	Uu	-

TABLE 3.9: Assigned labels for the second half of each combination of experimental categories. **B, I, S** stands for *baseline, indicative, subjunctive*, and **C, M, P** stands for *common, mixed, proper* The presence of a label/s or **?** means that there were at least 5 pairs in in which 3 raters agreed upon a label, the opposite situation is indicated with **-**. **?** means there was no majority vote for any label among the different pairs. **/** means that the votes were equally divided.

Matrix verbs, that is, predicates that embed another predicate or event, have their own veridicality value, or, in other words, they have an effect on the factuality value of the predicate they embed. According to this effect, matrix verbs are classified into three groups: implicative, factive and epistemic. The first group designates the matrices whose embedded events are a consequence of the whole complex event, like *manage* in *He managed to sell the house.* The second one denotes those whose embedded predicate

is a fact or a precondition for the whole complex predicate, like *know* in *He knew her father had arrived*. Lastly, the group of epistemic verbs includes those whose embedded event is a possibility, like *think* in *He thinks that Peter has arrived*. An extreme case of how this distinction affects the baseline of our annotations are examples 7 and 8, whose assigned labels are CT+ and PS+ respectively.

(7) a. La Razón sabía que la noticia
La Razón know.PST.IPFV.IND.3SG that the.F.SG news.F.SG
era falsa.
be.PST.IPFV.IND.3SG fake.PTCP.M.SG
La Razón knew that the news was fake.

b. La noticia era falsa.
the.F.SG news.F.SG be.PST.IPFV.IND.3SG fake.PTCP.M.SG
The news was fake.

(8) a. La junta directiva del Barça
the.F.SG board.F.SG governing.F.SG of.the.M.SG Barça
pensó que la reunión con los
think.PST.PFV.IND.3SG that the.F.SG meeting.F.SG with the.M.PL
abogados había ido bien.
lawyer.M.PL have.PST.IPFV.IND.3SG go.PTCP.M.SGwell
Barça's board of directors thought that the meeting with the lawyers had gone well.

b. La reunión con los abogados había
the.F.SG meeting.F.SG with the.M.PL lawyer.M.PL have.PST.IPFV.IND.3SG
ido bien.
go.PTCP.M.SGwell
The meeting with the lawyers had gone well.

In the first case, the matrix verb *saber* (to know) makes our hypothesis 7b a fact, or in other words, it assigns a high factuality value from our label scale, like *certainly yes*, to the pair. This assignation can surely be modified by other factors, but it definitely serves as a basic guideline to explain its behaviour with respect to the label of the pair below. In this second pair, the verb *pensar* (to think), is an epistemic verb, thus yielding 8b a possibility, and therefore assigning to it a more uncertain factuality value to begin with than for 7b. Furthermore, if we look at the other baseline variants from these pairs, we see more signs of this behaviour. In the case of 7, the other 2 variants have the same assigned label, CT+, whereas in the case 8, the other variants do not even have an assigned label. Given these differences, it will be prudent to consider their classification when defining the predictions and analyzing the data from the final study.

As to differences between the mood categories, there is not enough agreement to make clear conclusions, but some remarks can be pointed out. First of all, the distinction between the baseline and the mood alternation categories in the possibility condition is quite clear. All adverbial baseline combinations have a higher factuality value than the other two categories, but the distinction between indicative and subjunctive is not so clear. In the case of the negation condition, is more difficult to establish tendencies given the greater lack of agreement, but nevertheless, it seems that there is a difference between the three categories, altough it is not clear how it works.

Further observations can be made about tables 3.8 and 3.9. The first one is that, consistent with the model estimates, the specificity categories do not appear to affect the resulting labels, thus it might be helpful to remove them from the final study. Contrary to this, it seems that the specificity conditions affect the labels, particularly in the negation condition and even if it is to cause more disagreement. Lastly, there might be some signs of a label split as seen in the baseline of the mood condition, but it could be due to the above mentioned issue of the matrix verb, and therefore more data and a more thorough analysis is needed.

To sum up, in this section I have presented evidence that can partially explain the unexpected values seen in the baseline category, also we have seen how there are some distinctions between the different mood conditions and categories, even if for the latter their significante of their effect is not yet clear. Likewise, there are a few signs of variations between the specificity conditions but not the specificity categories. Lastly, we have also seen the overall need for a deeper linguistic analysis of the annotated pairs, analysis that I will present on the corpus of the final study.

3.6 Discussion and Conclusions

Throughout this chapter I have presented different results. Out of these one of the main results is the inter-annotator agreement, $AC_2 = 0.484$, that although higher than the initial chosen score, it is still lower than previous research on veridicality, with the work of de Marneffe et al. [1] having a $\kappa^f = 0.53$ and that of Saurí and Pustejovsky [5] having $\kappa^{cohen} = 0.81$, thus causes of this lower agreement need to be considered.

As already mentioned in Section 3.4, there are some factors in the experimental procedure that could have influenced the results. First of all, there was an important mistake: the labels were not presented in order. This could have prevented annotators from internalising the order behind the label set and thus made it more difficult for them to use it with ease. Second of all, the instructions might have not been clear enough given that at least a couple of annotators did not use their intuition as native Spanish speakers, but rather their basic formal linguistic knowledge; therefore I cannot say that these annotations were collected with a purely pragmatic approach but rather with a slightly mixed approach. Another reason why the instructions might not have been clear enough is that a couple of annotators reported the task to be quite difficult and at least one needed further clarifications. Given these important observations, the instructions for the final study were elaborated more thoroughly and more care was put into ensuring a correct experimental procedure.

An additional probable cause of the disagreement encountered might be the introduction of the label NaS. As shown in Section 3.5.1, even though this label was added as a mechanism to ensure the acceptability of the mood alternation pairs in the adverbal condition, given the high number of baseline pairs in which it was used and its very sparse distribution (no pair was labelled as *not a sentence*), this label was clearly misused. Considering this, and the fact that the NaS label does not completely fit into the scale shown in Figure 3.1, in the final study this label was not be brought into the analysis and was used as a filter.

Even though I should aim to improve the experimental design and procedure to increase agreement among annotators, it should be noted that the *real* agreement is likely not too high due to several factors. First of all, the approach used here is a pragmatic approach and, as already stated, a pragmatic approach means embracing uncertainty [1], which translates into lower agreement scores than in other settings. Second of all, as stated in Section 2.2.1, mood alternation is a phenomenon that reflects different presuppositions made about the information presented, and therefore, it is classified as a pragmatic phenomenon rather than a semantic one. Consequently, even more uncertainty and the consequential greater disagreement should be expected from the annotations. Lastly, as the work from de Marneffe et al. [1] and that of Pavlick and Kwiatkowski [9] demonstrate, there appears to be cases in which the *true* label is split, that is, that to represent the judgments of speakers not one, but two labels are required. In this study I have not seen strong evidence that supports this idea, but as stated in Section 3.5.5, more and better data might prove this.

Regarding the validity of the experimental design used for this pilot study there are some changes that can be made based on the effects seen and on the believe that a simplification of the design will clarify the results and therefore ease their understanding. The main change was to remove the specificity categories (common, mixed and proper) given that no significant effect from them was detected in the model coefficients and no signs of possible effects were seen in the labels assigned. Another change was to remove the possibility condition. Even though it has proved to have a significant effect in the labels chosen, there are, to my knowledge, more resources and data for the negation condition, and also the tendencies seen in tables 3.8 and 3.9 are more compelling. Concerning the specificity conditions (individual and collective), even though there is no evidence to prove that they have a significant effect in the annotations, there are some signs of an influence that, with a better experimental procedure might become a significant or close to significant effect. Lastly, it was noticed in Section 3.5.5 that there was an important linguistic variable that was left out when defining the experimental desing of this study, that of the matrix verbs. Since this factor can clearly influence the results, it must be included as a variable in the final study, but its exact definition will depend on the data gathered to build the corpus. With this we accomplish the second goal of this pilot study, ensuring the correctness of the experimental design and procedure.

As to the third goal, exploring different possibilities of statistical analysis, I can now define the main steps to be used in the final study. First, I will examine the overall distribution of labels and the distribution of labels per batches, as well as calculate counts like those in Table **??**, since all three have proven to be relevant to the posterior computations of inter-annotator agreement scores. As to these scores, the only one to be used in the final study is Gwet's AC_2 since it has proven to better reflect the annotations collected and its implementation is settled. Lastly, given how informative the regression model has proven to be, a cumulative link mixed model will be fit to the data with probably a more thorough analysis of its characteristics. Aside from all this statistical analysis and as stated in the previous section, it should not be forgotten that what here was a very concise linguistic analysis, it needs to be more detailed for the final study.

Finally, concerning the first and second research questions, that is, whether the negation and possibility conditions affect the factuality of the embedded event, when existing, the main event otherwise; based on the evidence presented here it can be said that they do affect the label assigned to the event. Contrary to this, there is not enough evidence to respond our third and fourth research questions affirmatively, that is, we cannot say that our individual and collective conditions have a significant effect on the label assigned. But given the already mentioned problems in the experimental procedure, these answers

should not be taken as final.

Next chapter presents the construction of the main or final study together with the analysis of the annotations gathered.

List of Figures

List of Tables

Bibliography

[1] Marie-Catherine de Marneffe, Christopher D Manning, and Christopher Potts. Did it happen? the pragmatic complexity of veridicality assessment. *Computational linguistics*, 38(2):301–333, 2012.

[2] Anastasia Giannakidou. (non) veridicality, evaluation, and event actualization: evidence from the subjunctive in relative clauses. In *Nonveridicality and Evaluation*, pages 17–49. Brill, 2014.

[3] Anastasia Giannakidou and Alda Mari. Mixed (non) veridicality and mood choice with emotive verbs. In *CLS 51*, 2015.

[4] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

[5] Roser Saurí and James Pustejovsky. Factbank: a corpus annotated with event factuality. *Language resources and evaluation*, 43(3):227–268, 2009.

[6] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*, 2018.

[9] Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, 2019.

[10] Alexis Ross and Ellie Pavlick. How well do nli models capture verb veridicality? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2230–2240, 2019.

[11] Aiala Rosá, Irene Castellón, Luis Chiruzzo, Hortensia Curell, Mathías Etcheverry, Ana Fernández Montraveta, Glòria Vázquez, and Dina Wonsever. Overview of fact at iberlef 2019: Factuality analysis and classification task. In *IberLEF@ SEPLN*, 2019.

[12] John Lyons. *Linguistic semantics: An introduction*. Cambridge University Press, 1995.

[13] David Sánchez-Jiménez. Una aproximación teórica a la definición del modo verbal español. 2011.

[14] RAE Real Academia Española. *Nueva gramática de la lengua española: Manual*. Espasa, 2011.

[15] José Manuel González Calvo. Sobre el modo verbal en español. *Anuario de estudios filológicos*, (18):177–204, 1995.

[16] Elisabeth Villalta. Mood and gradability: an investigation of the subjunctive mood in spanish. *Linguistics and philosophy*, 31(4):467–522, 2008.

[17] Errapel Mejías-Bikandi. Pragmatic presupposition and old information in the use of the subjunctive mood in spanish. *Hispania*, pages 941–948, 1998.

[18] Tris J Faulkner. *A Systematic Investigation of the Spanish Subjunctive: Mood Variation in Subjunctive Clauses*. Georgetown University, 2021.

[19] Ingrid Falk and Fabienne Martin. Towards an inferential lexicon of event selecting predicates for french. *arXiv preprint arXiv:1710.01095*, 2017.

[20] María Amparo Alcina Caudet. *Las expresiones referenciales. Estudio semántico del sintagma nominal*. PhD thesis, Universitat de València, 1999.

[21] Fernando García Murga. *Las presuposiciones lingüísticas*. Servicio Editorial de la Universidad del País Vasco/Euskal Herriko ..., 1998.

[22] Lucille Herrasti et al. *Características semánticas definitorias de la presuposición*. PhD thesis, El Colegio de México, 2011.

[23] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[24] Patrick E Shrout. Measurement reliability and agreement in psychiatry. *Statistical methods in medical research*, 7(3):301–317, 1998.

[25] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[26] Kenneth J Berry, Janis E Johnston, and Paul W Mielke Jr. Weighted kappa for multiple raters. *Perceptual and motor skills*, 107(3):837–848, 2008.

[27] Kerrie P Nelson and Don Edwards. Measures of agreement between many raters for ordinal classifications. *Statistics in medicine*, 34(23):3116–3132, 2015.

[28] Amalia Vanacore and Maria Sole Pellegrino. Robustness of κ-type coefficients for clinical agreement. *Statistics in Medicine*, 41(11):1986–2004, 2022.

[29] Matthijs J Warrens. Equivalences of weighted kappas for multiple raters. *Statistical Methodology*, 9(3):407–422, 2012.

[30] Matthijs J Warrens. Corrected zegers-ten berge coefficients are special cases of cohen's weighted kappa. *Journal of Classification*, 31(2):179–193, 2014.

[31] Anthony J Conger. Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, 88(2):322, 1980.

[32] Robert L Brennan and Dale J Prediger. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and psychological measurement*, 41(3):687–699, 1981.

[33] Kilem Li Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61 (1):29–48, 2008.

[34] KL Gwet. Irrcac: Computing chance-correcte agreement coefficients (cac) version 1.0, 2019.

[35] Rune Haubo B Christensen. Cumulative link models for ordinal regression with the r package ordinal. *Submitted in J. Stat. Software*, 35, 2018.